A Lightweight SAR Image Ship Detection Method Based on Improved Convolution and YOLOv7

Hongdou Tang , Song Gao * , Song Li , Pengyu Wang , Jiqiu Liu , Simin Wang , Jiang Qian

Posted Date: 23 October 2023

doi: 10.20944/preprints202310.1446.v1

Keywords: synthetic aperture radar (SAR); lightweight networks; ship detection; YOLOv7

Article

A Lightweight SAR Image Ship Detection Method Based on Improved Convolution and YOLOv7

Hongdou Tang [1], Song Gao [1,*], Song Li [1], Pengyu Wang [1], Jiqiu Liu [1], Simin Wang [1] and Jiang Qian [2]

[1] The College of Mechanical and Electrical Engineering, Chengdu University of Technology, Chengdu 610059, China
[2] The School of Resources and Environment, University of Electronic Science and Technology of China, Chengdu 611731, China
* Correspondence: gs@cdut.edu.cn

Abstract: The airborne and satellite-based synthetic aperture radar enables the acquisition of high-resolution SAR oceanographic images in which even the outlines of ships can be identified. The detection of ship targets from SAR images has a wide range of applications, such as the military, where the dynamic grasp of enemy targets can help improve the early warning capability of naval defence, and the civilian detection of illegal fishing vessels can help improve the level of maritime management. Due to the density of ships in SAR images, the extreme imbalance between foreground and background clutter, and the diversity of target sizes, achieving lightweight and highly accurate multi-scale ship target detection remains a great challenge. To this end, this paper proposes an attention mechanism for multiscale receptive fields convolution block (AMMRF). AMMRF not only makes full use of the location information of the feature map to accurately capture the regions in the feature map that are useful for detection results, but also effectively captures the relationship between the feature map channels, so as to better learn the relationship between the ship and the background. Based on this, a new YOLOv7-based ship target detection method,You Only Look Once SAR Ship Identification (YOLO-SARSI), is proposed, which acquires the abstract semantic information extracted from the high-level convolution while retaining the detailed semantic information extracted from the low-level convolution. Compared to the deep learning detection methods proposed by previous authors, our model is more lightweight, only 18.43 M. We examined the effectiveness of our method on two SAR image public datasets: the High-Resolution SAR Images Dataset (HRSID) and the Large-Scale SAR Ship Detection Dataset-v1.0 (LS-SSDD-V1.0). The results show that the average accuracy (AP_{50}) of the detection method YOLO-SARSI proposed in this paper on the HRSID and LS-SSDD-V1.0 datasets is 4.9% and 5% higher than that of YOLOv7, respectively.

Keywords: synthetic aperture radar (SAR); lightweight networks; ship detection; YOLOv7

1. Introduction

Synthetic Aperture Radar (SAR), with its all-weather, all-day, weather-independent imaging characteristics, has become one of the most important tools for terrestrial observation. It transmits electromagnetic pulses to the target area through an antenna, receives electromagnetic pulses back from the target area, compares the received and transmitted electromagnetic pulses, and generates images by the Doppler effect.SAR operates in an electromagnetic waveband that penetrates clouds and dust, which allows it to provide remote sensing images in complex weather environments. With airborne and satellite-based SAR, it is possible to obtain high-resolution SAR images of the ocean, and ship targets as well as the ship's tracks are clearly visible in these images. Therefore, ship detection systems using SAR have been widely used in maritime surveillance activities and play an increasingly important role [1–3]. Among the ship target detection methods, Constant False Alarm Rate (CFAR) is one of the classical algorithms widely used for ship target detection, which detects ship targets by modelling the statistical distribution of background clutter [4].This traditional algorithm is

 doi:10.20944/preprints202310.1446.v1

suitable for SAR images with simple backgrounds and does not process well in images with complex backgrounds. In 2012, AlexNet, proposed by Alex Krizhevsky et al. [5] made a splash in the ImageNet image recognition competition, crushing the classification performance of the second place support vector machines (SVM). After this, convolutional neural networks (CNNs) have received renewed attention. It has been easy to encounter the problem of gradient disappearance in CNNs. In 2015, Kaiming He et al. proposed ResNet [6], a network with a residual block that alleviates the gradient disappearance and has had a profound impact on the design of subsequent deep neural networks. With the development of deep learning, current deep learning algorithms have far surpassed the performance of traditional machine learning algorithms. Applying deep learning to image processing can significantly improve detection accuracy and speed for tasks such as target detection and instance segmentation [5]. Currently, many authors have applied deep learning to SAR ship target detection. Kang et al. [7] proposed a multilayer fusion convolutional neural network based on contextual regions and verified that contextual information has an impact on the performance of the neural network in recognising ships in SAR images. Jiao et al. [8] fused features of different resolutions through dense connections for solving the multi-scale and multi-scene SAR ship detection problem. Cui et al. [9] integrated feature pyramids with convolutional block attention modules to integrate salient features with global unambiguous features to improve the accuracy of ship detection in SAR images. To improve the detection speed of SAR image ship, Zhang et al .[10] proposed a high-speed ship detection method for SAR images based on grid convolutional neural network (G-CNN). Qu et al. [11] proposed an anchor freed detection model based on mask-guided features to reduce computational resources and improve the performance of ship detection in SAR images. Sun et al. [12] proposed a model based on a densely connected deep neural network with an attention mechanism (Dense-YOLOv4-CBAM) to enhance the transmission of image features. Liu et al. [13] based on YOLOv4, through feature pyramid network (FPN) [6] to obtain multi-scale semantic information and use scale-equalizing pyramid convolution (SEPC) to balance the correlation of multi-scale semantic information, and proposed SAR-Net. Wang et al. [14] added multi-scale convolution and transformer module to YOLO-X to improve the performance of YOLO-X in detecting ships. In the FBR-Net network proposed by Fu et al. [15], the designed ABP structure uses a layer-based attention approach and a spatial attention approach to balance the semantic information of the features in each layer, making the network more focused on small ships. In order to improve the detection of small ships in complex background SAR images, Guo et al. [16] combined feature refinement, feature fusion and head enhancement methods to design a high-precision detector called CenterNet++. Considering that contextual information is crucial for the detection of small and dense ships, Zhao et al. [17] proposed a new CNN-based method in which as many small ships as possible are first proposed and then combined with contextual information to exclude spurious ships from the predictions, improving the accuracy of ship detection in SAR images.

All of the above researches have contributed to the improvement of the accuracy of ship target detection in SAR images, but the following problems still exist:

1. Most of their SAR image ship target detection frameworks are designed for small target ships in SAR images, and in the process of designing, the performance of recognizing multi-scale ships is not well considered.Therefore, the detection accuracy decreases for the presence of multi-scale ships in the SAR image.

2. Some networks use complex feature fusion in the Neck part, and it is the fusion of features extracted from high level convolutions of the backbone network, while the semantic details about the ship extracted from low level convolutions are easy to be “drowned out” due to the stacking of the convolutions, which is not friendly to ship detection.

3. These methods are mainly dedicated to the improvement of the detection accuracy of ship targets in SAR images, Sun et al. used DenseNet in each layer of convolution in the backbone network,Wang et al. added Transformer on YOLOX,This will definitely increase the number of parameters in the model, without taking into account the reduction of redundant parameters and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 October 2023 doi:10.20944/preprints202310.1446.v1

computational costs.The redundancy in the feature maps of convolutional neural networks leads to a large consumption of memory and computational resources [18].

To address the above problems, we propose a new detection framework based on YOLOv7 [19]. The new detection method is more lightweight and works well for multi-scale ship target detection in SAR images.Our contributions can be summarized as follows.

1. A new convolutional block, which we name AMMRF, is proposed. for SAR images containing ships, it obtains feature information from different sensory fields and filters this feature information, making the network more focused on information useful for ship detection.

2. The addition of AMMRF to the backbone network of YOLOv7 makes the backbone network more dexterous.The addition of AMMRF makes the whole detection framework complete with feature fusion in the backbone network. Therefore, we modified the Neck part of YOLOv7 by removing the complex feature fusion. We named the new detection framework as YOLO-SARSI.

3. The number of parameters in YOLO-SARSI is very small, only 18.43M, which is 16.36M less compared to YOLOv7. Even so, the accuracy of YOLO-SARSI in SAR images of ship targets is still higher than that of YOLOv7.

2. Methods

2.1. Analysis of SAR Image Features

The mainstream object detection frameworks proposed in the past, such as YOLO series, Fast-CNN [20], etc., use common objects in context (COCO) [21] or the PASCAL visual object classes (PASCAL VOC) [22] dataset to measure the performance of the recognition framework.Both the COCO dataset and the PASCAL VOC dataset are widely available object detection databases with a rich set of objects, containing 80 classes of objects in the COCO dataset and 20 classes of objects in the PASCAL VOC dataset.

Optical color images are image data acquired by visible and partially infrared band sensors and will usually contain grayscale information in multiple bands to facilitate target identification and classification extraction. Figure 1 illustrates some typical pure background SAR images and optical images of their corresponding regions.

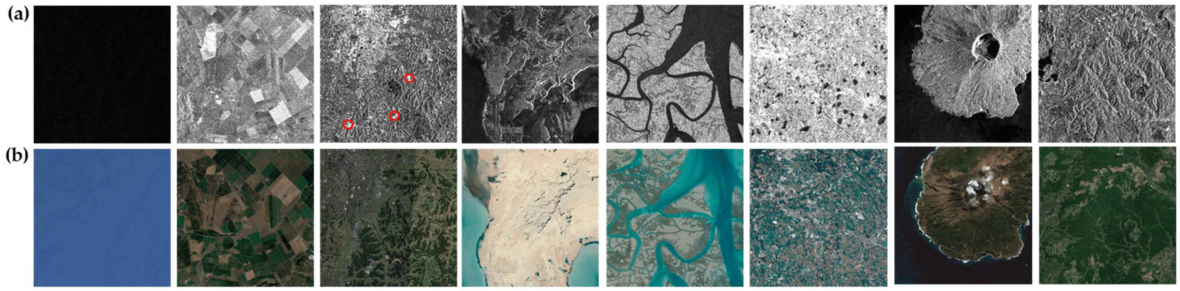


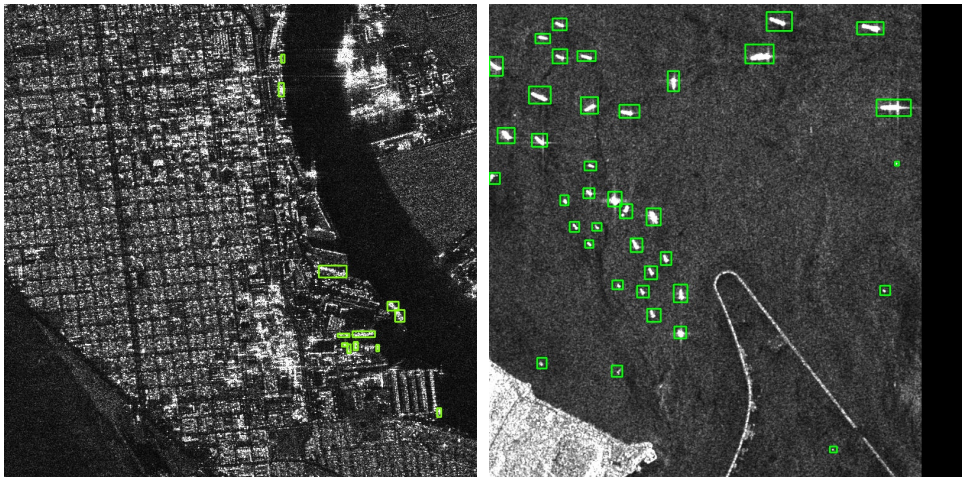
Figure 1. Abundant pure backgrounds of SAR images in literature [23]. (**a**) SAR images; (**b**) optical images. Sea surface; farmlands; urban areas; Gobi; remote rivers; villages; volcanos; forests.

Unlike the optical color images in the COCO dataset and the PASCAL VOC dataset, the high-resolution SAR images dataset [24] (HRSID) and large-scale SAR ship detection dataset-v1.0 [23] (LS-SSDD-V1.0) are grayscale images, which record only the echo information of one electromagnetic wave band, and the pixel points on the image are the reflections of the ground target to the radar wave, and the value of each pixel in the image is a sample, which only represents the energy of the electromagnetic wave reflected by the ground target received by SAR.This also leads to the fact that SAR images themselves do not carry as much semantic information as optical color images.In the radar system, rough ground targets have a higher backscatter of the radar's transmitted electromagnetic

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 October 2023 doi:10.20944/preprints202310.1446.v1

waves, which means that more of the reflected electromagnetic waves can be picked up by the radar. Smooth ground targets generally have almost no return signal and the radar will only receive high reflected electromagnetic waves when the radar beam is perpendicular to the surface of such features. If most of the reflected electromagnetic waves from a ground target are returned to the SAR, the target will appear as a bright area in the SAR image, and vice versa as a dark area. Thus, flat and smooth targets often appear as dark areas in the SAR image and rough targets appear as bright areas in the SAR image. Objects of metallic, high dielectric constant materials, where the polarisation direction of the incident wave is not necessarily parallel to the length direction of the target, but as long as there is an electric field component parallel to it, it will produce resonance effect, forming a strong echo. In the HRSID and LS-SSDD-V1.0 datasets, ships are often shown as bright blocks or bright spots, water areas often behave as dark areas, and land areas are mostly bright areas.

There are a large number of small ships in both datasets. In Figure 2 the small ships are small in pixel size, carry less semantic information and have fewer discriminative features.



(a) (b)

Figure 2. Ships marked in green boxes in a complicated ocean background. Figure (**a**) shows the SAR image from the HRSID dataset and Figure (**b**) shows the SAR image from LS-SSDD-V1.0.

Next, we analyse the process of human identification of ships in SAR images. Firstly the global information of the image is acquired and the areas of sea, harbour and sea and river banks are identified. Secondly, we acquire the local information of the image to determine the bright spots or highlights in the sea, harbour and river banks, and obtain more detailed information about these bright spots or highlights to determine whether they are ships. Some of the ships in the SAR image retain the ship's shape, while others are simply bright blocks or bright spots. It is easy for human to identify ships in the sea by picking up bright spots or highlights in the sea. In contrast, harbours and riverbanks have very poor visual effects and are difficult to identify, so human needs to obtain more detailed information about the ships in these areas, such as the brightness of the ship, whether it has the shape of a ship and the relationship between the surrounding pixels. When human identifies a certain type of target in an image, they can easily understand the relationships between the image global and image local and between image localities, and unconsciously use the information reflected in these relationships when identifying such targets in that image. As can be seen, when human identifies a ship in a SAR image, extracting information about the image global and local is essential for identifying the ship.

From the above analysis, it is clear that:

1. SAR images are grey-scale images, and the images carry little information. Complex and excellent detection frameworks are not necessarily suitable for SAR image ship target detection, and there may be redundancy of convolution when using these detection frameworks in recognising ships.
2. There are many small ships in the SAR images, and the smaller ships carry less

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 October 2023 doi:10.20944/preprints202310.1446.v1

semantic information, which can easily be confused with other interference, leading to missed or wrong detection.

3. As can be seen from human approach to ship detection, the network model requires global information about the image as well as high quality semantic detail information about the ship itself.

Inspired by this, we consider that the low-level convolutional blocks should always retain the semantic information of the ship itself, while the global information of the image can be extracted through the superposition of the convolutional blocks.The entire recognition model should be as lightweight as possible, which means minimizing the number of convolutions in the model.

2.2. Improved Convolution Block:AMMRF

Our proposed convolution block is shown in Figure 3, named attention mechanisms for multiscale receptive fields convolution block (AMMRF) for the convenience of exposition. It can be divided into three parts: 3×3 convolution for extracting features, 1×1 convolution mainly for reducing the number of feature map channels, and coordinate attention block [25] (CA) for enhancing the ability of the convolution block to learn feature representation.

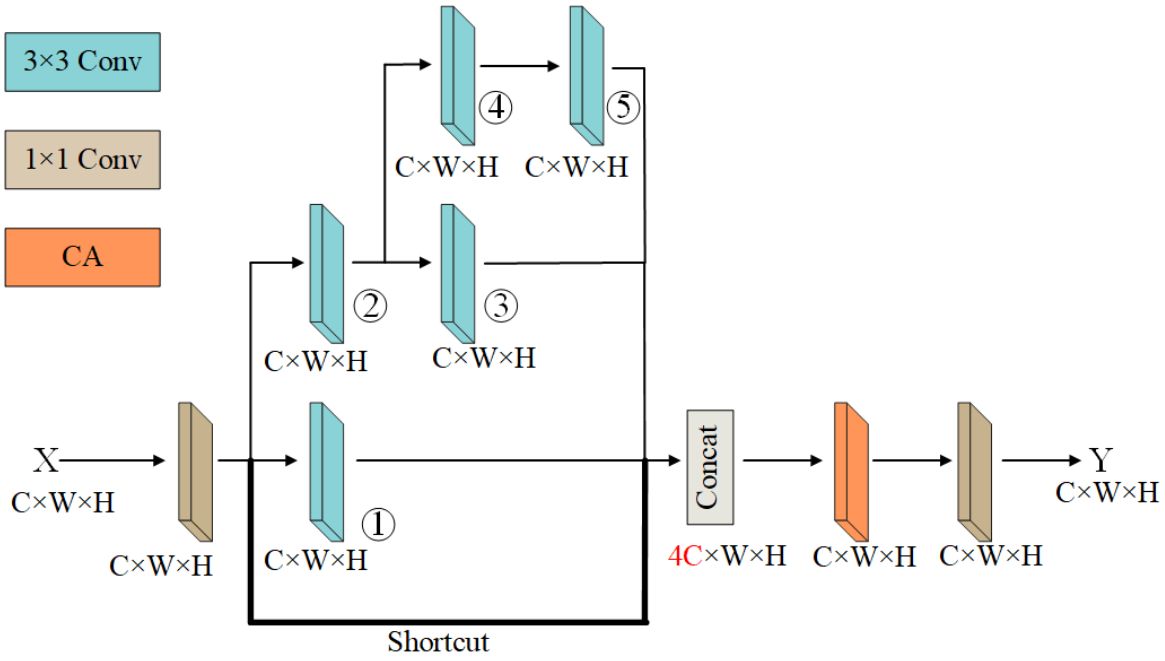


Figure 3. The overall structure of SAR Dectection Convolution (AMMRF). Concat's form of concatenation allows to obtain a feature map with four times the number of channels as the input feature map X.

GoogLeNet [26] made a big splash in the ImageNet competition in 2014, where the inception block used to extract information from different spatial dimensions of the image by convolution of different sizes, thus allowing feature information to be extracted on different sensory fields. Figure 3. labels the five 3×3 convolutions in the AMMRF block as 1 to 5, which can extract feature information on different receptive fields. Convolution 1 has a receptive field of 3×3 on the input feature map. The stacking of convolutions 2 and 3 makes them have a receptive field of 5×5 on the input feature map. The stacking of convolutions 2, 4 and 5 makes them have a receptive field of 7×7 on the input feature map.

The residual structure of ResNet alleviates the problem of gradient disappearance to a certain extent, while the feature information extracted from the low-level convolution can be retained and output to the high-level convolution. The semantic information carried by the ship itself in the SAR image is relatively small, and the lack of feature information in the detection process can easily be confused with other interference, leading to missed and wrong detection and affecting the final results.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 October 2023 doi:10.20944/preprints202310.1446.v1

In order to enable the lower layer convolution to extract the semantic information of the ship itself to the higher layer, we introduced the shortcut connection in ResNet in the AMMRF. This enhances the information flow between the front and back layers, and the feature map information on the input side is also retained on the output side, which makes the weak information of the ship itself less likely to be overwhelmed, and mitigates the gradient disappearance, making the network training faster.

ResNet uses a summation method to sum up the feature maps in the channel direction, which results in a loss of dimensionality and feature information. DenseNet [27] uses a stitching method to superimpose all the feature maps in the channel direction, which can retain the feature information better than ResNet. Therefore, the output of the five 3×3 convolutions and the output of the shortcut concatenation was obtained using the Concat concatenation form of DenseNet. This form of concatenation superimposes different feature maps on the channels, enabling the fusion of features in the channel dimension, mapping the features to the interaction space, and better learning the relationship between the ship and the background.

Concat's form of concatenation allows to obtain a feature map with four times the number of channels as the input feature map X. To capture the relationships between these channels, Concat is followed by CA. CA not only makes full use of the captured location information so that the region of interest can be captured accurately, it is also effective in capturing the relationships between channels, which effectively enhances the ability of the AMMRF to learn feature representation.

In order to reduce the number of operations and parameters, we have reduced the number of channels in the feature map by using 1×1 convolution at the input and output.

2.3. Network Structure

In this subsection, the differences between our network structure and YOLOv7 will be compared and then the advantages of YOLO-SARSI will be described.The network architecture of YOLOv7 [19] is shown in Figure 4. YOLOv7 is an anchor based detection framework. The input image is fed into the backbone to extract features, and the backbone consists of Stem, ELAN and MP1; the feature maps extracted from the backbone are processed by head to output three layers of feature maps with different sizes, and the neck consists of SPPCSPC, UpSample, ELAN-H, REP and MP2. Finally, the output is processed by detect prediction results.

Based on YOLOv7, we propose a new network architecture as shown in Figure 5, which we named "You Only Look Once - SAR image Ship Identification" (YOLO-SARSI). It is also divided into three parts: the backbone for feature extraction, the neck for reprocessing and rationalizing the features extracted from the backbone network, and the head for final prediction detection.

In the backbone, we replaced the ELAN of YOLOv7 with AMMRF, while removing one layer of MP1 and ELAN. When stacking AMMRF to extract features from SAR images, the extracted feature maps always accept feature information of different convolutional layer depths, and the feature maps obtained by deep convolution always retain the details of the images extracted by shallow convolution feature information. Therefore, in the backbone, we use the stacking of AMMRF, so that the features extracted from the backbone incorporate information from different scales and different depths of convolution, and the presence of residual links in AMMRF can make the deep convolution also retain the features proposed by the shallow convolution. This means that the feature information extracted in the deep layer network retains both global and local information of the image, as well as fine-grained feature information of the image, such as the semantic information of the ship itself. YOLO-SARSI is still an anchor based single stage target detection model. In the backbone of YOLO-SARSI, four 3×3 convolutions form the Stem block, which extracts the detailed information of the image itself and reduces the size of the output to one quarter of the input, completing the downsampling operation, which reduces the number of parameters of the model.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 October 2023 doi:10.20944/preprints202310.1446.v1

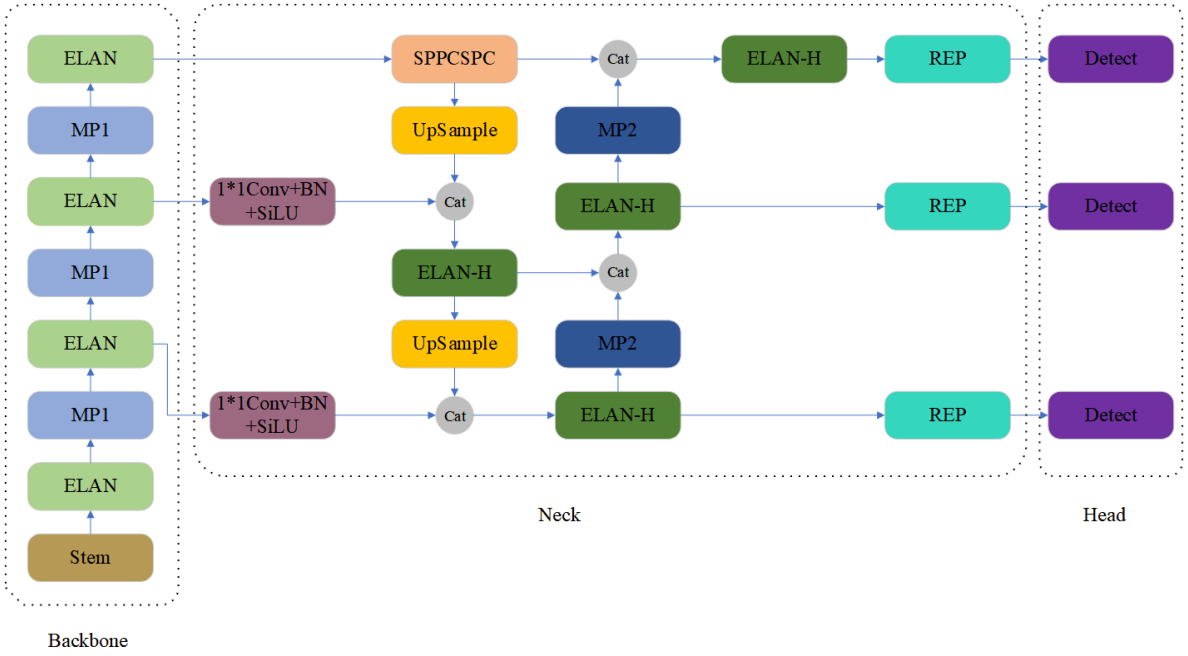


Figure 4. The overall structure of YOLOv7. The overall structure of YOLOv7 is drawn from the code provided by the authors of YOLOv7.

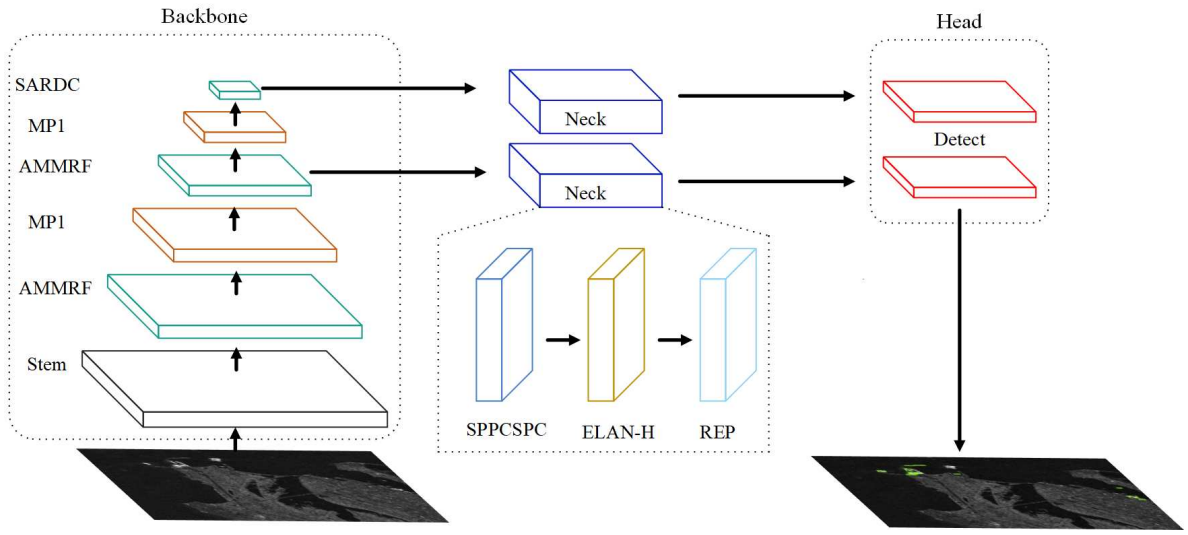


Figure 5. The overall structure of YOLO-SARSI.

In the neck, we still use the same convolutional block as in YOLOv7, except that we do not use any feature fusion in this part, and the number of feature maps output from the backbone network to the neck is reduced from three to two. SPPCSPC, ELAN-H, REP and MP1 all use the blocks in YOLOv7. In YOLOv7, upsampling was used to fuse the small size features from the high level convolution to the large size features from the low level convolution, but upsampling often has some side effects, such as noise amplification. If downsampling is used to fuse the feature map obtained from the lower convolution to the small size feature map obtained from the higher convolution, there will be redundancy of features. Therefore, in the YOLO-SARSI neck, we do not use any feature fusion. This also allows the model to have a smaller number of parameters.

In the anchor based YOLO series, the feature maps extracted from three different depths of convolutional layers of the backbone network are used to identify targets. The feature maps extracted from these three different convolutional layers have different sizes and are used to detect targets of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 October 2023 doi:10.20944/preprints202310.1446.v1

three different sizes. In the two datasets HRSID and LS-SSDD-V1.0, the proportion of large targets is very small. In YOLO-SARSI, the output feature maps of the second AMMRF layer and the third AMMRF layer are used to detect small ships and large and medium-sized ships respectively.

3. Experimental Results

In this paper, all experiments were done on a cloud server equipped with an NVIDIA V100-SXM2 (32G graphics memory) graphics processing unit (GPU). We used the Python 3.8 compiled language to implement the training and CUDA 11.3 to accelerate the computations. In the experiments, the SDD300 [28], Cascade R-CNN [29] , Faster R-CNN [30] , Mask R-CNN [31] are all based on the mmdetection platform [32] , YOLOv7 is derived from the publicly available source code by the authors of YOLOv7.

There are 5604 cropped SAR images and 16951 ships in HRSID and 9000 cropped SAR images and 6015 ships in LS-SSDD-V1.0. The LS-SSDD-V1.0 dataset has more pure background images. The image size in both datasets is 800 x 800 $pixels^2$. The ship pixel area in the image is used to measure the ship size, i.e., the relative pixel size, rather than the physical size. The average ship pixel area in LS-SSDD-v1.0 is only 381 $pixels^2$, while the average ship pixel area in HRSID is 1808 $pixels^2$. The dataset is divided into three types of ships based on their pixel area size: small ships (pixel area less than 482 $pixels^2$), medium ships (pixel area between 482 - 1452 $pixels^2$) and large ships (pixel area greater than 1452 $pixels^2$) [23,24]. The statistics of the number of ships of three sizes in the two datasets are shown in Figure 6.

The number of iterations for the training cycle on the HRSID dataset was 60. The LS-SSDD-V1.0 dataset had smaller ship targets and more complex images, so the number of iterations for the training cycle on the LS-SSDD-v1.0 dataset was 128. The dataset was provided with an image size of 800 $\times$ 800 and the input sizes in the recognition framework were all 800 $\times$ 800.

No pre-trained models were used for any of the training. That is, all models were trained from scratch. In this paper, a model is considered successful in detecting a ship target when the IOU value between its prediction frame and the real target frame is higher than 0.5. In order to accurately evaluate the detection performance of each model, the MS COCO evaluation metric and Parameters metric were used in this paper.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 October 2023 doi:10.20944/preprints202310.1446.v1

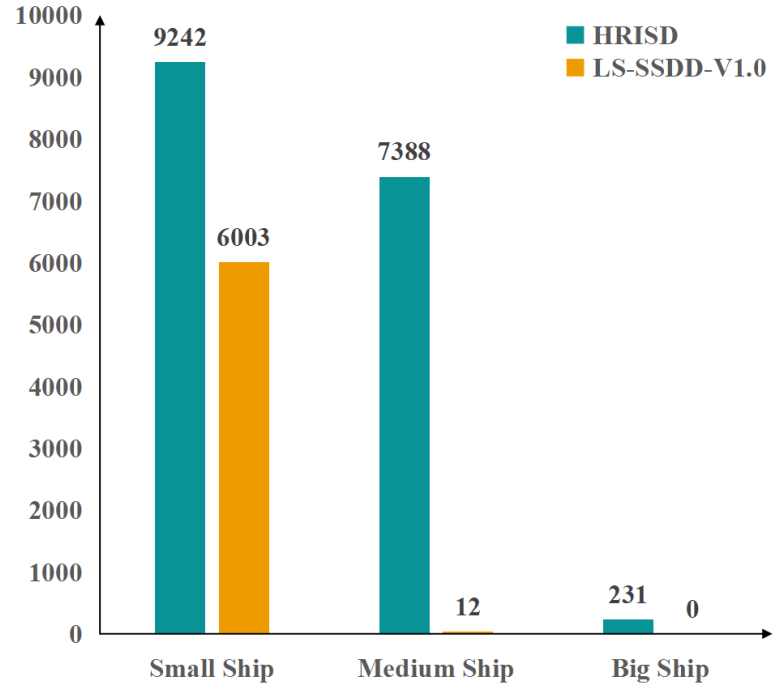


Figure 6. Comparison of the number of ships of three sizes in the two datasets.

3.1. YOLO-SARSI Recognition Accuracy Evaluation

We tested some of the other good algorithms on the HRSID dataset and the LS-SSDD-V1.0 dataset for comparison with our algorithm, as Table 1 is shown. It can be seen that YOLO-SARSI has a significant advantage over the excellent two stage detection algorithms Cascade R-CNN, Faster R-CNN and Mask R-CNN, both in terms of AP_{50}, and the number of parameters of the model. Compared with YOLOv7, YOLO-SARSI improved 4.9 % on AP_{50} and 3.3 % on $AP_{50:95}$ for the HRSID dataset. Most of the ships in LS-SSDD-V1.0 are small targets, which have less information on themselves and are difficult to detect, and although they only improved 1.3 % on $AP_{50:95}$ for this dataset, they improved significantly by 5.0 % on AP_{50}. YOLO-SARSI has fewer network parameters, which requires less hardware storage space when deployed in an embedded chip.Although the number of parameters in SDD300 is only 5.32M more than the model presented in this paper, the AP_{50} and $AP_{50:95}$ of YOLO-SARSI are much higher on both datasets.YOLO-SARSI is not only a lightweight model, but also a high-precision model.

Table 1. Experimental results for the datasets.

Dataset	Model	AP_{50}(%)	$AP_{50:95}$(%)	Params(M)
HRSID	Cascade R-CNN	65.1	41.2	68.93M
	Faster R-CNN	69.8	43.6	41.12
	Mask R-CNN	69.9	43.9	41.12
	SDD 300	56.5	36.8	23.75
	YOLOv7	80.9	59.6	34.79
	YOLO-SARSI	**85.8**	**62.9**	**18.43**
LS-SSDD-V1.0	Cascade R-CNN	55.4	20.1	68.93M
	Faster R-CNN	63.4	23.9	41.12
	Mask R-CNN	63.3	24.1	41.12
	SDD 300	32.5	10.1	23.75
	YOLOv7	61.6	25.0	34.79
	YOLO-SARSI	**66.6**	**26.3**	**18.43**

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 October 2023 doi:10.20944/preprints202310.1446.v1

There are 2396 more images in the LS-SSDD-V1.0 dataset than in the HRSID dataset, but the detection results of the same model on the LS-SSDD-V1.0 dataset are not as good as those on the HRSID dataset. Compared to the HRSID dataset, the image quality of the LS-SSDD-V1.0 dataset was worse. The ships in the LS-SSDD-V1.0 dataset are basically small targets, which makes it very difficult for the model to identify the features that the ships themselves carry, such as the shape of the ship, from these small targets.

Meanwhile, there are only 0.67 ships per image on average in the LS-SSDD-V1.0 dataset, while there are 3.02 ships per image in the HRSID dataset, which is 4.51 times more than the former. In the LS-SSDD-V1.0 dataset, the small number of ship targets tends to cause an imbalance between positive and negative samples of the data, which tends to lead to a large number of negative samples making the training process ineffective and the loss gradient of negative samples tends to dominate, leading to a decrease in the performance of the model.

Based on the experimental results, we plotted the precision recall curves (PR curve) of each model on the HRISD dataset and LS-SSDD-V1.0 dataset, as shown in Figure 7.

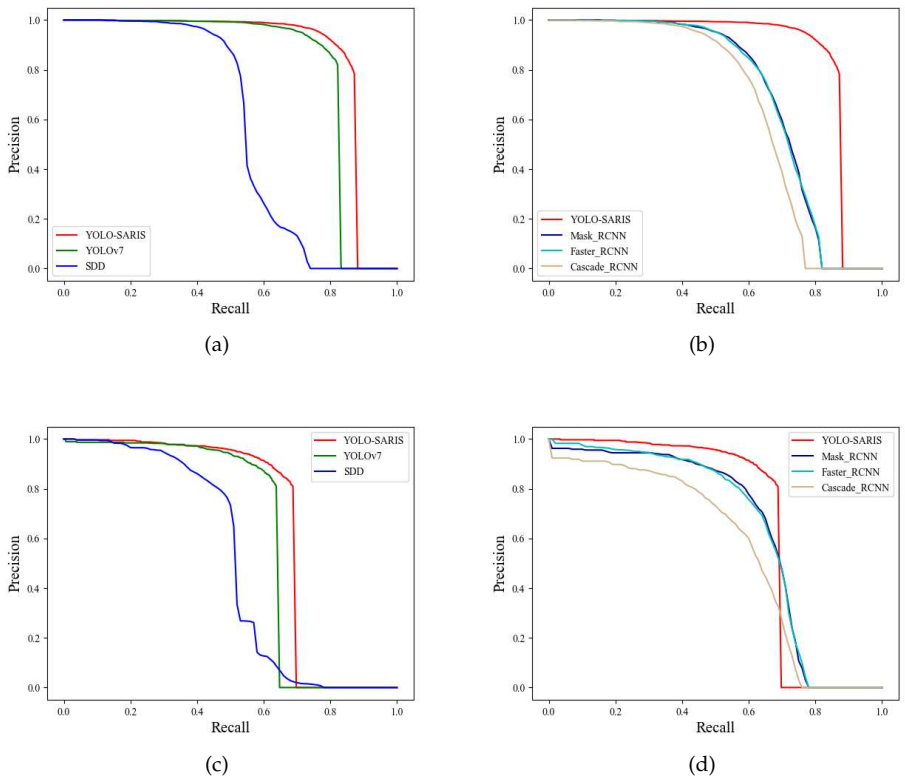


Figure 7. PR curve of different methods. (**a**): One stage models with HRSID; (**b**): Two stage models with HRSID; (**c**): One stage models with LS-SSDD-V1.0; (**d**): Two stage models with LS-SSDD-V1.0;

The horizontal axis of the PR curve curve is recall and the vertical axis is precision, which reflects the relationship between precision and recall. The area between the curve and the two axes is the AP_{50} . The higher the recall and precision, the better the model, i.e., the more convex the PR curve is, the better the model. The more convex the PR curve is, the better the model is. If the PR curve of one model is completely surrounded by the curve of another model, it can be concluded that the latter is better than the former. From Figure 7 , it can be seen that the PR curve of YOLO-SARSI encompasses all other models. These results show that YOLO-SARSI's detection performance on both datasets is significantly better than YOLOv7 and the other models.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 October 2023 doi:10.20944/preprints202310.1446.v1

3.2. Instance Testing

In order to see the performance of YOLO-SARSI on specific images, we selected three images from the HRISD dataset and LS-SSDD-V1.0 dataset, respectively, for inference in YOLOv7 and YOLO-SARSI, and the inference results are shown in Figures 8 and 9 are shown.

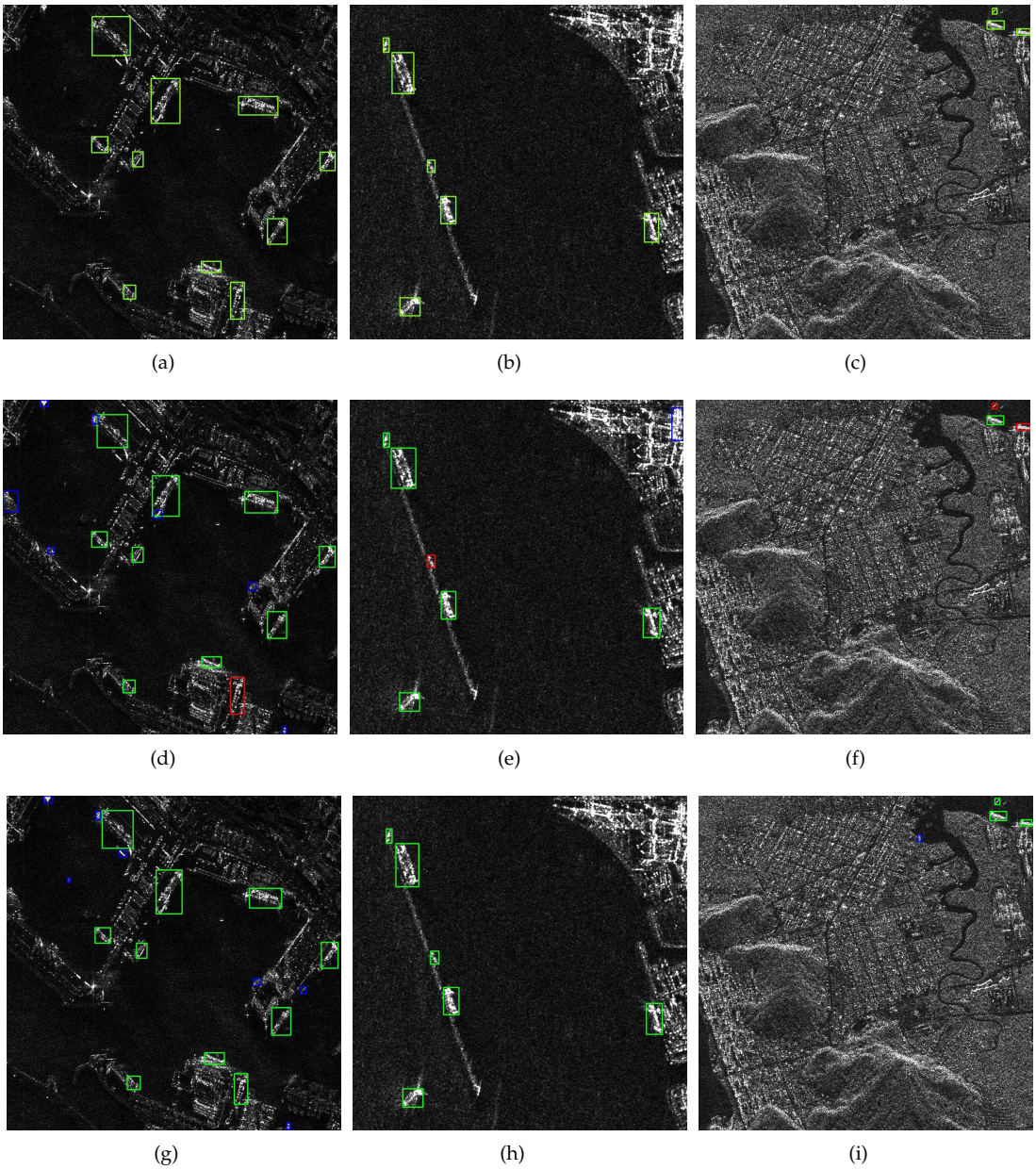


Figure 8. Diagram of HRISD results. Green marked boxes are ships that were detected correctly or true marked boxes, blue marked boxes are ships that were detected incorrectly and red marked boxes are ships that were missed.(**a**–**c**):ground truth; (**d**–**f**):YOLOv7; (**g**–**i**): YOLO-SARSI.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 October 2023 doi:10.20944/preprints202310.1446.v1

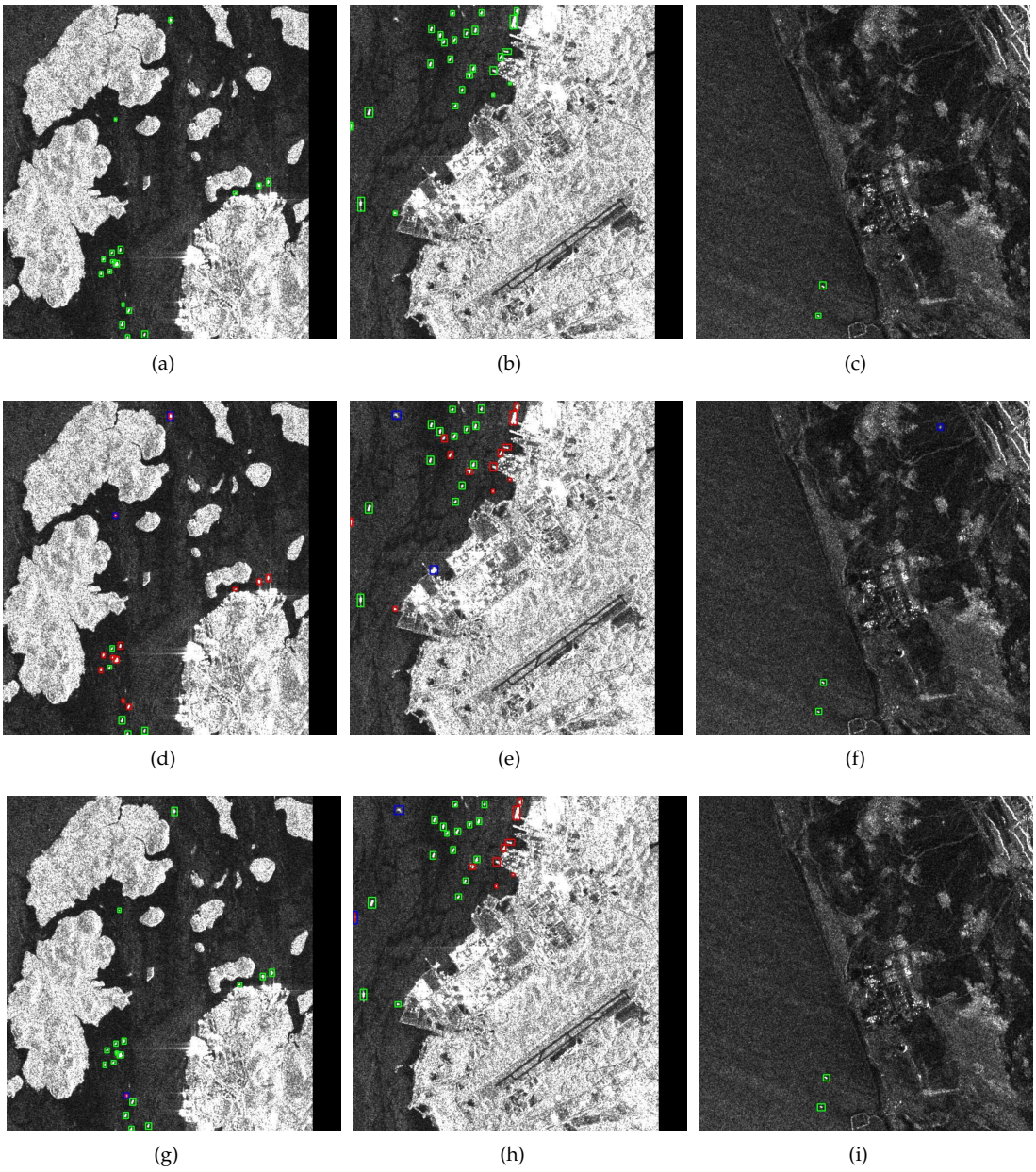


Figure 9. Schematic of LS-SSDD-V1.0 results. Green marked boxes are ships that were detected correctly or true marked boxes, blue marked boxes are ships that were detected incorrectly and red marked boxes are ships that were missed.(**a**–**c**): ground truth; (**d**–**f**): YOLOv7; (**g**–**i**): YOLO-SARSI.

In Figure 8a, the area of land and water are almost the same, and the ships are close to the shore with a complex background. In Figure 8b, the ships are mainly in the water, but some of them have trailing noise. In Figure 8c, the water is mainly in the upper right and lower left corners, mostly land, and only three ships are in the upper right corner of the image. YOLOv7 has the phenomenon of identifying non-ship objects as ships in Figure 8a, while there are missed detections, and although YOLO-SARSI has false detections, there are no missed detections. In Figure 8b YOLOv7 has missed and false detections, while YOLO-SARSI perfectly detects all the ships. The recognition accuracy of YOLOv7 in Figure 8c is only 33.3%. Although YOLO-SARSI also has a false detection in this image, it is not difficult to find that the target of the false detection is the prominent color spot on the shore, and YOLO-SARSI detects all the ships in Figure 8c.

In Figure 9a there are many islands, some of which are even as large as the ship in the image.

 doi:10.20944/preprints202310.1446.v1

The ships in Figure 9b are mainly distributed in the water, but there are a large number of ships on shore, which are difficult to identify because the ships are easily confused by background clutter. The land and water in Figure 9c have similar grayscale, while there are some bright spots on the land.YOLO-SARSI perfectly detected all the ships in Figure 9a,c. In Figure 9b, although there are missed ships, the ships they missed are basically shore-based ships.

In order to find out what features the model has learned, whether the features it has learned are what we expect, or whether the model has learned cheating information, a heat map visualization of the model's gradient calculation results in the image is performed.

Gradient-weighted class activation mapping (Grad-CAM) [33] can help us to analyse the areas of focus of a network model on a particular class of targets, so that the areas of focus of the network can in turn be used to analyse whether the network has learned the correct information or features.Gradient information of the feature maps obtained from the second layer of anchor in YOLOv7 and YOLO-SARSI for the three images selected from the HRSID dataset were plotted using Grad-CAM, as shown in Figure 10.

Figure 10d–i, the color depth of the pixel points, reflects the information of the area where the model focuses on the image. The brighter the color, the more the model focuses on the feature information at this location, which means that more information about the ship is extracted at this location as well.In Figure 10b,c, YOLOv7 is concerned with a lot of information that is not related to the ship.Combined with the real annotation frame, the feature gradient map of YOLO-SARSI is more vivid in color compared to YOLOv7 for a real ship at the same location in an image, which also shows that YOLO-SARSI can better focus on the pixel information of the ship itself.

Figure 11 plots the gradient information of the feature maps obtained from the first layer of anchor in YOLOv7 and YOLO-SARSI for the three images in the LS-SSDD-V1.0 dataset using Grad-CAM. It can be seen that YOLOv7 focuses on a more haphazard information for the images, which is particularly evident in Figure 11f. It even focuses a lot on objects in the land, which leads to it having a false detection. YOLO-SARSI focuses a lot on the ships, giving a lot of attention to the ships in the images.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 October 2023 doi:10.20944/preprints202310.1446.v1

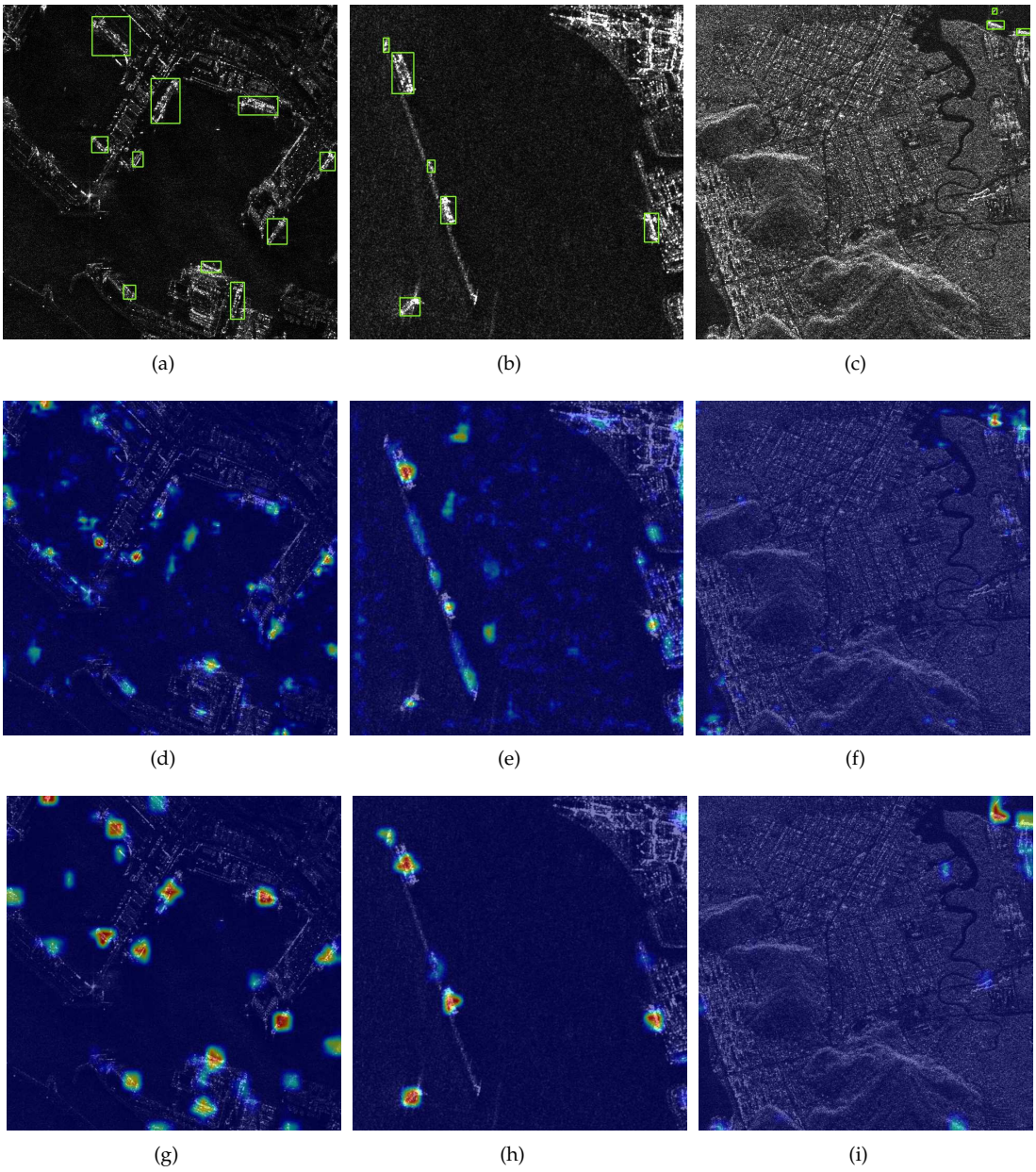


Figure 10. Schematic heat map of HRSID results. (**a**–**c**): ground truth; (**d**–**f**): YOLOv7; (**g**–**i**): YOLO-SARSI.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 October 2023 doi:10.20944/preprints202310.1446.v1

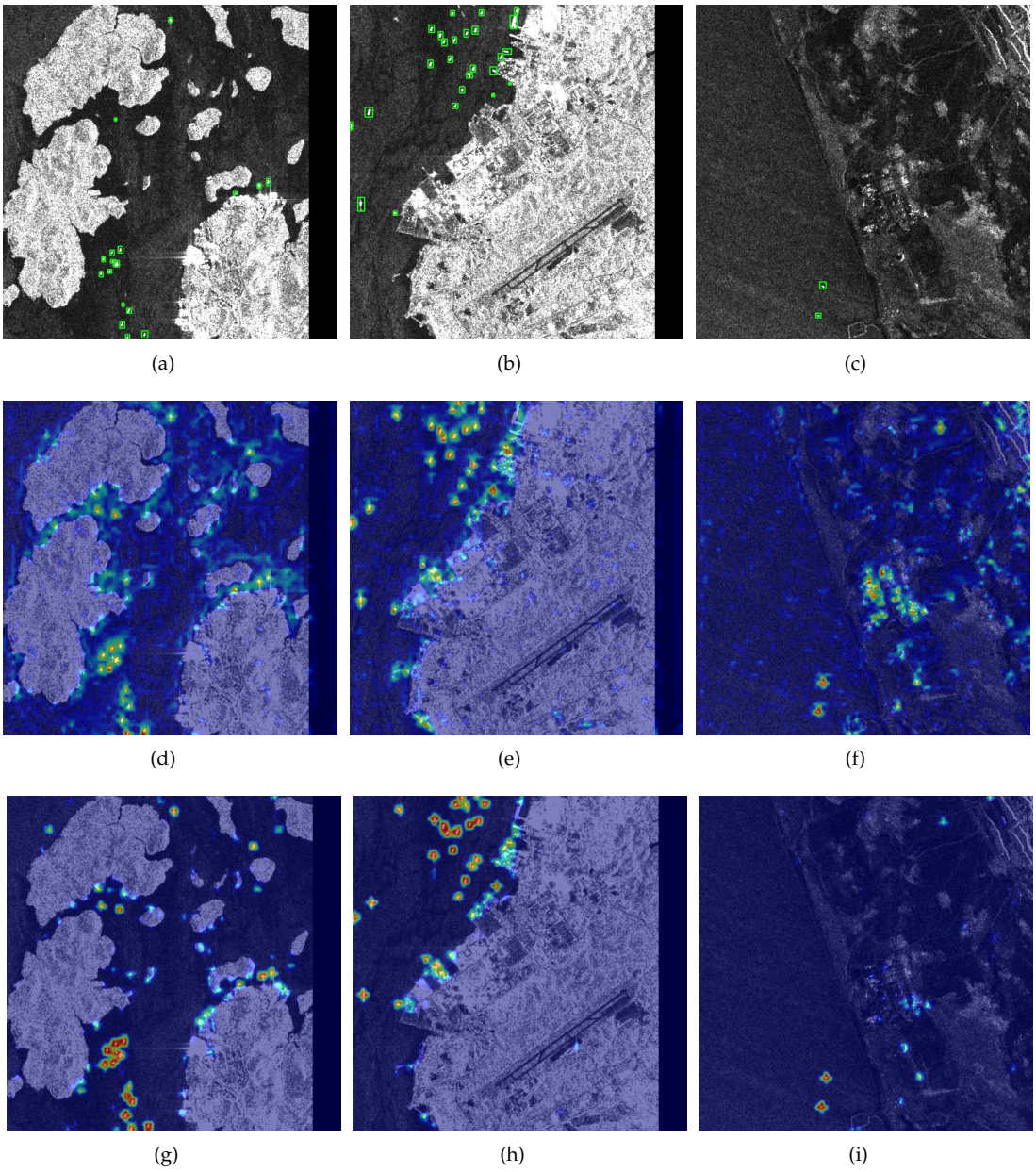


Figure 11. Schematic heat map of LS-SSDD-V1.0 results. (**a**–**c**): ground truth; (**d**–**f**): YOLOv7; (**g**–**i**): YOLO-SARSI.

4. Conclusions

In this paper, we design a new convolutional block AMMRF, starting from dissecting the difference between SAR images and optical colour images, and analysing the way and basis for human identification of SAR images.Based on this, we propose a new network model for ship detection of SAR images based on YOLOv7, which we name YOLO-SARSI. The network model is used in HRISD and LS-SSDD- V1.0, two publicly available datasets, the results show that YOLO-SARSI has a good performance in terms of average accuracy and model size metrics. The lightweight YOLO-SARSI means that our models can be more easily integrated into embedded systems.It is hoped that this paper can provide some guidance for developers and researchers exploring the field of SAR ship detection to obtain better detection performance in practical industrial applications.

Author Contributions: Conceptualization, H.T.; Data curation, H.T.; Formal analysis, H.T.; Funding acquisition, S.G.; Investigation, H.T. and S.L.; Methodology, H.T. and S.G.; Resources, H.T. and S.L.; Software, H.T. and Simin

Wang; Supervision, S.G. and S.L.; Validation, H.T., PW., J.L. and S.W.; Writing – original draft, Hongdou Tang and S.G.; Writing – review & editing, H.T., SG., J.L., S.W. and J.Q.

Funding: This work was supported by the National Natural Science Foundation of China (NSFC) under Grant No. 41930112.

References

1. Dai, H.; Du, L.; Wang, Y.; Wang, Z. A Modified CFAR Algorithm Based on Object Proposals for Ship Target Detection in SAR Images. *IEEE Geoscience and Remote Sensing Letters* **2016**, *13*, 1925–1929. https://doi.org/10.1109/LGRS.2016.2618604.
2. Tang, T.; Xiang, D.; Xie, H. Multiscale salient region detection and salient map generation for synthetic aperture radar image. *Journal of Applied Remote Sensing* **2014**, *8*.
3. Xiang, D.; Tang, T.; Ni, W.; Zhang, H.; Lei, W. Saliency Map Generation for SAR Images with Bayes Theory and Heterogeneous Clutter Model. *Remote Sensing* **2017**, *9*. https://doi.org/10.3390/rs9121290.
4. Hwang, S.I.; Ouchi, K. On a Novel Approach Using MLCC and CFAR for the Improvement of Ship Detection by Synthetic Aperture Radar. *IEEE Geoscience and Remote Sensing Letters* **2010**, *7*, 391–395. https://doi.org/10.1109/LGRS.2009.2037341.
5. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks **2017**. *60*. https://doi.org/10.1145/3065386.
6. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *CoRR* **2015**, *abs/1512.03385*, http://xxx.lanl.gov/abs/1512.03385.
7. Kang, M.; Ji, K.; Leng, X.; Lin, Z. Contextual Region-Based Convolutional Neural Network with Multilayer Fusion for SAR Ship Detection. *Remote Sensing* **2017**, *9*. https://doi.org/10.3390/rs9080860.
8. Jiao, J.; Zhang, Y.; Sun, H.; Yang, X.; Gao, X.; Hong, W.; Fu, K.; Sun, X. A Densely Connected End-to-End Neural Network for Multiscale and Multiscene SAR Ship Detection. *IEEE Access* **2018**, *6*, 20881–20892. https://doi.org/10.1109/ACCESS.2018.2825376.
9. Cui, Z.; Li, Q.; Cao, Z.; Liu, N. Dense Attention Pyramid Networks for Multi-Scale Ship Detection in SAR Images. *IEEE Transactions on Geoscience and Remote Sensing* **2019**, *57*, 8983–8997. https://doi.org/10.1109/TGRS.2019.2923988.
10. Zhang, T.; Zhang, X. High-Speed Ship Detection in SAR Images Based on a Grid Convolutional Neural Network. *Remote Sensing* **2019**, *11*. https://doi.org/10.3390/rs11101206.
11. Qu, H.; Shen, L.; Guo, W.; Wang, J. Ships Detection in SAR Images Based on Anchor-Free Model With Mask Guidance Features. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2022**, *15*, 666–675. https://doi.org/10.1109/JSTARS.2021.3137390.
12. Sun, B.; Wang, X.; Li, H.; Dong, F.; Wang, Y. Small-Target Ship Detection in SAR Images Based on Densely Connected Deep Neural Network with Attention in Complex Scenes. *Applied Intelligence* **2022**, *53*, 4162–4179. https://doi.org/10.1007/s10489-022-03683-1.
13. Gao, S.; Liu, J.M.; Miao, Y.H.; He, Z.J. A High-Effective Implementation of Ship Detector for SAR Images. *IEEE Geoscience and Remote Sensing Letters* **2022**, *19*, 1–5. https://doi.org/10.1109/LGRS.2021.3115121.
14. Wang, S.; Gao, S.; Zhou, L.; Liu, R.; Zhang, H.; Liu, J.; Jia, Y.; Qian, J. YOLO-SD: Small Ship Detection in SAR Images by Multi-Scale Convolution and Feature Transformer Module. *Remote Sensing* **2022**, *14*. https://doi.org/10.3390/rs14205268.
15. Fu, J.; Sun, X.; Wang, Z.; Fu, K. An Anchor-Free Method Based on Feature Balancing and Refinement Network for Multiscale Ship Detection in SAR Images. *IEEE Transactions on Geoscience and Remote Sensing* **2021**, *59*, 1331–1344. https://doi.org/10.1109/TGRS.2020.3005151.
16. Guo, H.; Yang, X.; Wang, N.; Gao, X. A CenterNet++ model for ship detection in SAR images. *Pattern Recognition* **2021**, *112*, 107787. https://doi.org/https://doi.org/10.1016/j.patcog.2020.107787.
17. Zhao, J.; Guo, W.; Zhang, Z.; Yu, W. A coupled convolutional neural network for small and densely clustered ship detection in SAR images. *Science China Information Sciences* **2018**, *62*, 1–16.
18. Qiu, J.; Chen, C.; Liu, S.; Zeng, B. SlimConv: Reducing Channel Redundancy in Convolutional Neural Networks by Weights Flipping. *CoRR* **2020**, *abs/2003.07469*, http://xxx.lanl.gov/abs/2003.07469.
19. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, 2022, http://xxx.lanl.gov/abs/2207.02696.

20. Girshick, R. Fast R-CNN. 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1440–1448. https://doi.org/10.1109/ICCV.2015.169.
21. Lin, T.Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C.L.; Dollár, P. Microsoft COCO: Common Objects in Context, 2015, http://xxx.lanl.gov/abs/1405.0312.
22. Everingham, M.; Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vision* **2010**, *88*, 303–338. https://doi.org/10.1007/s11263-009-0275-4.
23. Zhang, T.; Zhang, X.; Ke, X.; Zhan, X.; Shi, J.; Wei, S.; Pan, D.; Li, J.; Su, H.; Zhou, Y.; Kumar, D. LS-SSDD-v1.0: A Deep Learning Dataset Dedicated to Small Ship Detection from Large-Scale Sentinel-1 SAR Images. *Remote Sensing* **2020**, *12*. https://doi.org/10.3390/rs12182997.
24. Wei, S.; Zeng, X.; Qu, Q.; Wang, M.; Su, H.; Shi, J. HRSID: A High-Resolution SAR Images Dataset for Ship Detection and Instance Segmentation. *IEEE Access* **2020**, *8*, 120234–120254. https://doi.org/10.1109/ACCESS.2020.3005861.
25. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design, 2021, http://xxx.lanl.gov/abs/2103.02907.
26. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions, 2014, http://xxx.lanl.gov/abs/1409.4842.
27. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2261–2269. https://doi.org/10.1109/CVPR.2017.243.
28. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In *Computer Vision – ECCV 2016*; Springer International Publishing, 2016; pp. 21–37. https://doi.org/10.1007/978-3-319-46448-0_2.
29. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving Into High Quality Object Detection. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 6154–6162. https://doi.org/10.1109/CVPR.2018.00644.
30. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, 2016, http://xxx.lanl.gov/abs/1506.01497.
31. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN, 2018, http://xxx.lanl.gov/abs/1703.06870.
32. Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J.; Zhang, Z.; Cheng, D.; Zhu, C.; Cheng, T.; Zhao, Q.; Li, B.; Lu, X.; Zhu, R.; Wu, Y.; Dai, J.; Wang, J.; Shi, J.; Ouyang, W.; Loy, C.C.; Lin, D. MMDetection: Open MMLab Detection Toolbox and Benchmark, 2019, http://xxx.lanl.gov/abs/1906.07155.
33. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision* **2019**, *128*, 336–359. https://doi.org/10.1007/s11263-019-01228-7.